Visualisation interactive et réexpression des données avec SAS/Insight

Monique Le Guen, Sophie Destandau, Dominique Ladiray

Visualisation interactive et réexpression des données avec SAS/Insight

La première partie du menu *Analyze* de SAS/Insight propose sept types de graphiques différents. La façon dont se distribue une variable sera représentée sous forme d'un *histogramme* ou d'un *Box Plot* si la variable en question est une variable d'intervalle, par un *Bar Chart* ou un *Mosaic Plot* s'il s'agit d'une variable nominale. Le *Line Plot* et le *Scatter Plot* (diagramme de dispersion, ou nuage de points) permettent de visualiser la façon dont se répartissent les observations suivant deux variables distinctes, le *Line Plot* étant plus particulièrement adapté à la représentation d'une série temporelle. Enfin, la façon dont se répartissent les observations suivant trois variables distinctes peut être visualisée au moyen d'un *Rotating Plot* (diagramme de rotation).

À l'exception des *Line Plots*, la totalité des graphiques présentés dans cet article ont été réalisés à partir d'une même table SAS incluant 173 observations, en l'occurrence 173 pays, et 25 variables, en particulier les taux de natalité (NAT) et de mortalité (MORT) pour 1 000 habitants, le taux d'accroissement naturel annuel de la population (ACCR) en %, le taux de fécondité ou nombre d'enfants par femme (FERTI), la part (en % de la population totale) des moins de 15 ans (AGE15) et des plus de 65 ans (AGE65), le produit national brut (PNB) et le taux d'urbanisation (URBA) en % de la population totale.

Histogramme et Box Plot

L'histogramme révèle la forme, ou plutôt une forme, de la distribution étudiée, dans l'exemple ci-dessous celle du taux de mortalité (variable MORT).

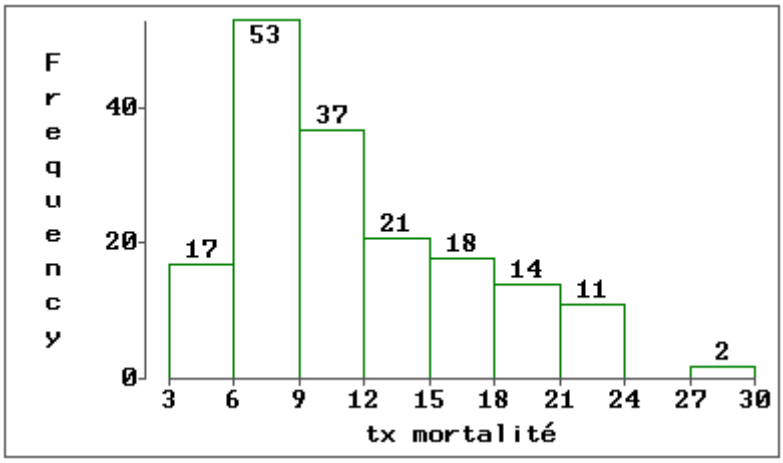
Le Box Plot, ici présenté en mode horizontal, apporte de nombreuses informations supplémentaires :

- la longueur de la boîte (rectangle central) détermine l'étendue de la partie centrale de la distribution (taux de mortalité compris entre les quartiles Q1 et Q3, valeurs 7 et 15) ;
- la bande verticale matérialisée à l'intérieur de la boîte indique la position de la médiane (valeur 10) ;
- l'étendue des queues de distribution hors points atypiques est déterminée par la longueur des moustaches (rectangles latéraux), étant précisé que l'extrémité de la moustache de gauche est ici[1] fixée par la plus petite valeur (= 3) supérieure ou égale à Q1 - 1,5 (Q3 - Q1), celle de la moustache de droite par la plus grande valeur (= 23) inférieure ou égale à Q3 + 1,5 (Q3 - Q1) ;
- de part et d'autre des moustaches sont mises en évidence les observations atypiques, avec valeur inférieure à Q1 - 1,5 (Q3 -Q1) ou supérieure à Q3 + 1,5 (Q3 - Q1), ici la Gambie et la Sierra Leone, où le taux de mortalité est particulièrement élevé ;
- la petite diagonale du losange superposé à la boîte indique la position de la moyenne (valeur 11,21) ;
- sa grande diagonale, de longueur 2 σ, permet d'apprécier la valeur de l'écart-type.

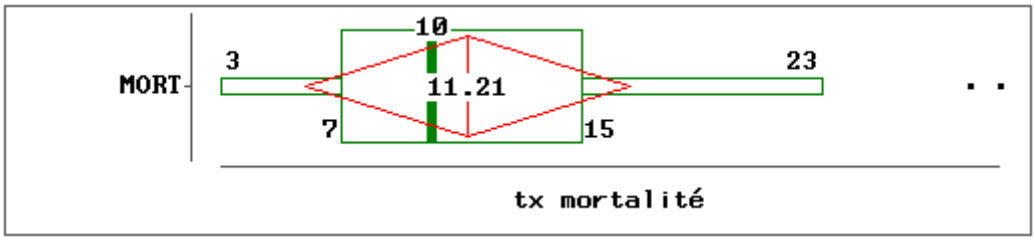
---

1. Le coefficient appliqué à (Q3 - Q1) est paramétrable.

Histogramme

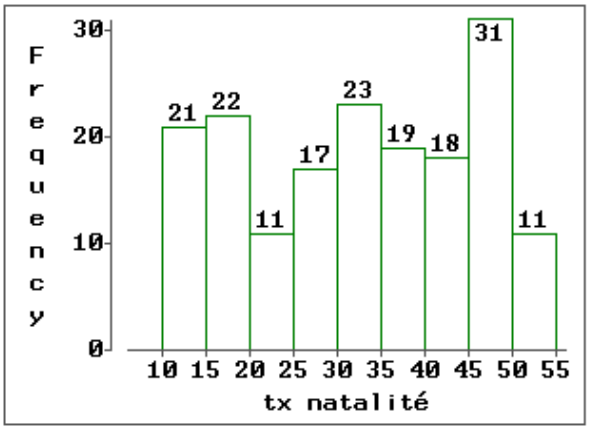


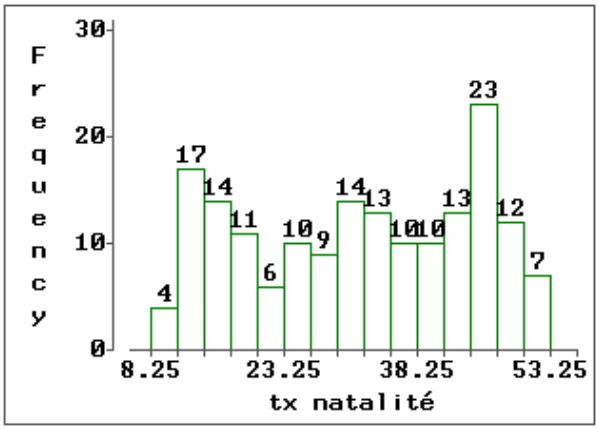
Box Plot



Redécoupage d'un histogramme

Avec SAS/Insight, il est très simple et très rapide, grâce à la souris, de modifier le découpage en classes de la variable étudiée. L'effet sur l'allure de l'histogramme peut être surprenant, comme le montre l'exemple ci-dessous relatif à la distribution du taux de natalité (pour 1 000 habitants).

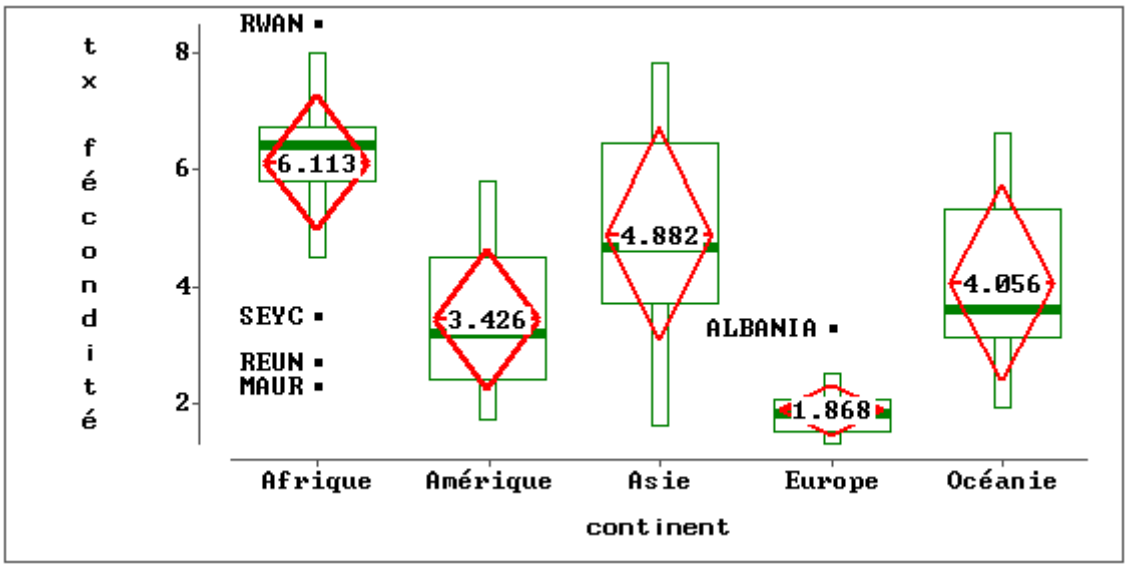




Juxtaposition de Box Plots

Les 173 pays observés ont été répartis en 5 groupes, selon le continent. Les 5 Box Plots juxtaposés ci-dessous, ici en mode vertical, représentent la distribution du taux de fécondité (nombre d'enfants par femme) dans les différents continents.

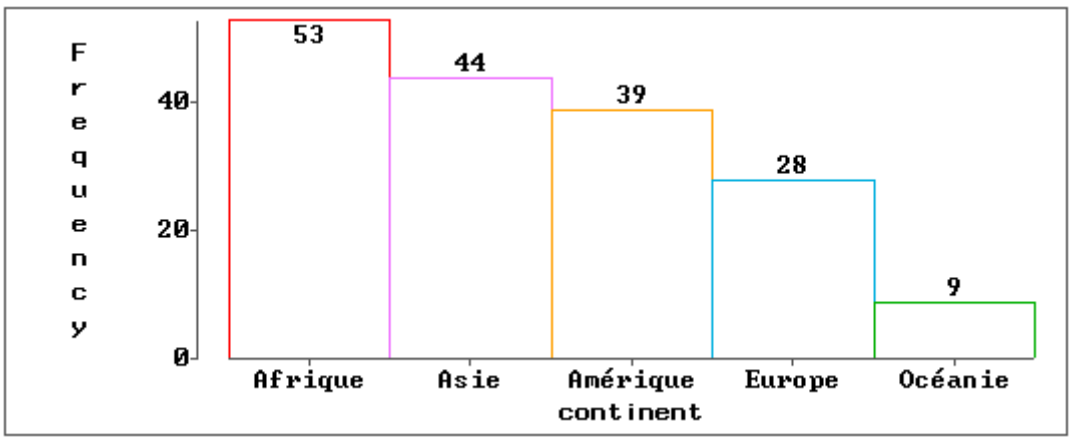
Particulièrement efficace, ce type de présentation permet d'apprécier en un coup d'oeil la façon dont se distribue une variable d'intervalle en fonction des modalités d'une variable nominale. En outre, il constitue une excellente introduction visuelle à l'analyse de la variance.



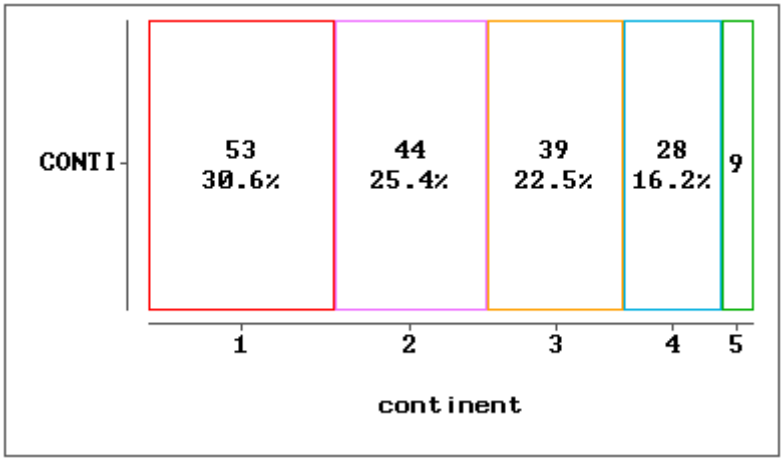
Bar Chart et Mosaic Plot

La façon dont se répartissent les observations selon les valeurs d'une variable nominale, ici la répartition des pays suivant le continent, peut être représentée au moyen d'un Bar Chart ou d'un Mosaic Plot.

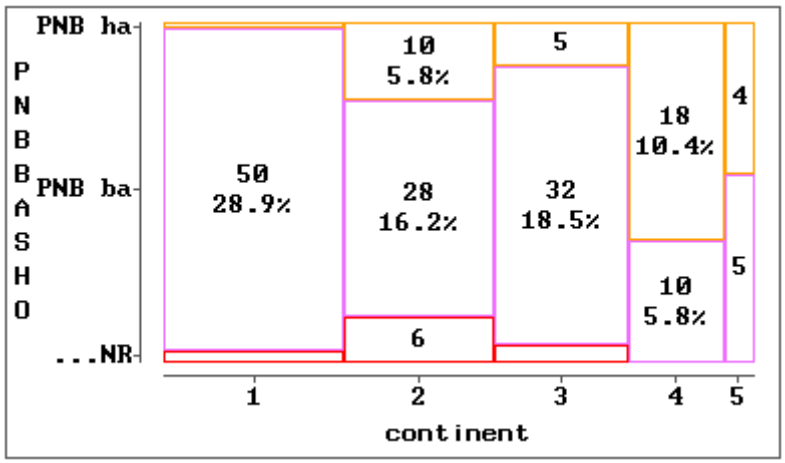
Bar Chart



Mosaic Plot



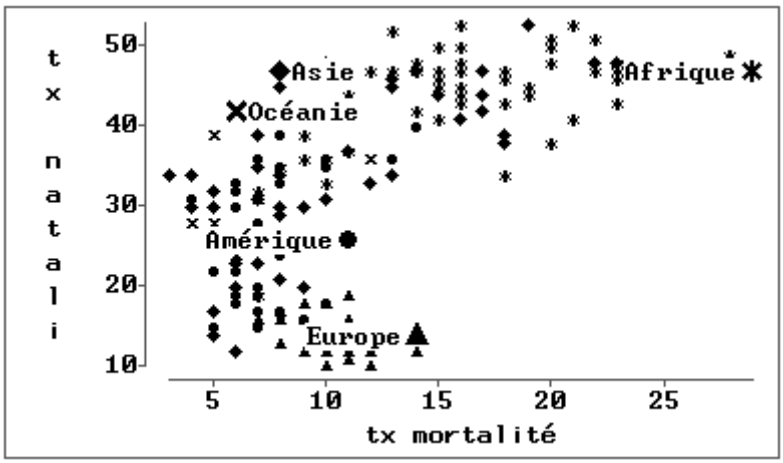
Une autre utilisation possible du Mosaic Plot, équivalent visuel de la procédure *Proc Freq* de SAS, est la représentation des croisements entre deux variables nominales. Ainsi, le graphique ci-dessous permet de visualiser la façon dont se répartissent les 173 pays observés selon le continent et le PNB (grossièrement découpé en 3 modalités : haut, bas, non-réponse).



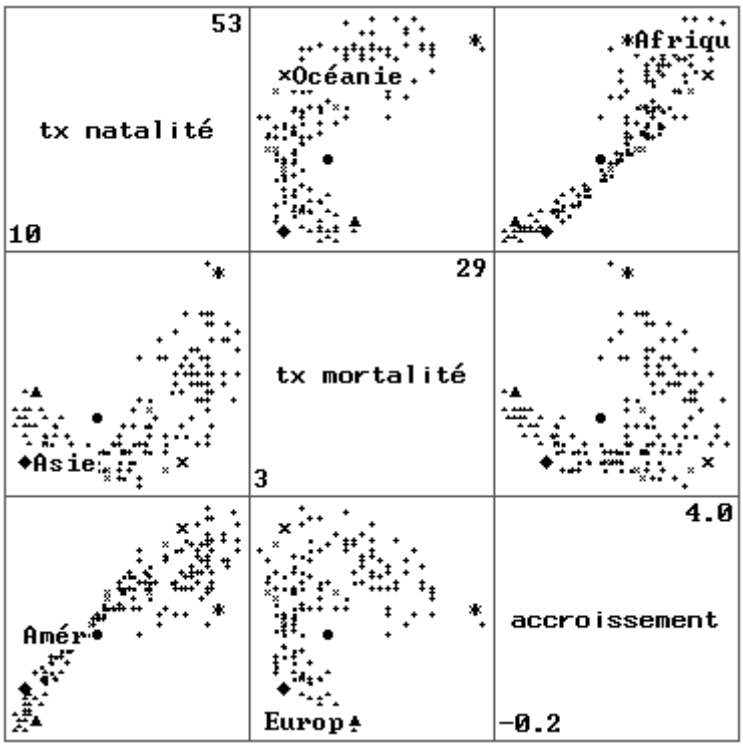
Le Scatter Plot

Le Scatter Plot, ou diagramme de dispersion, permet tout à la fois d'apprécier le type de liaison (linéaire ou autre) pouvant exister entre deux variables, de diagnostiquer une éventuelle hétéroscédasticité, de repérer les groupes (*clusters*) et les observations atypiques (*outliers*).

Le Scatter Plot ci-dessous visualise la façon dont se répartissent les 173 pays observés suivant le taux de natalité et le taux de mortalité. Y a été ajoutée une troisième dimension, via l'utilisation de cinq marqueurs différents repérant le continent d'appartenance des pays (on peut ainsi remarquer une forte concentration de pays africains dans la partie supérieure droite du diagramme). On aurait tout aussi bien pu en en ajouter une quatrième, repérage par exemple des pays selon la tranche de PNB, au moyen de la couleur.



Les liaisons 2 à 2 entre plusieurs variables d'intervalle, ici les taux de natalité et de mortalité (pour 1 000 habitants) et le taux d'accroissement naturel de la population (en %), peuvent être visualisées au moyen de matrices de diagrammes de dispersion.



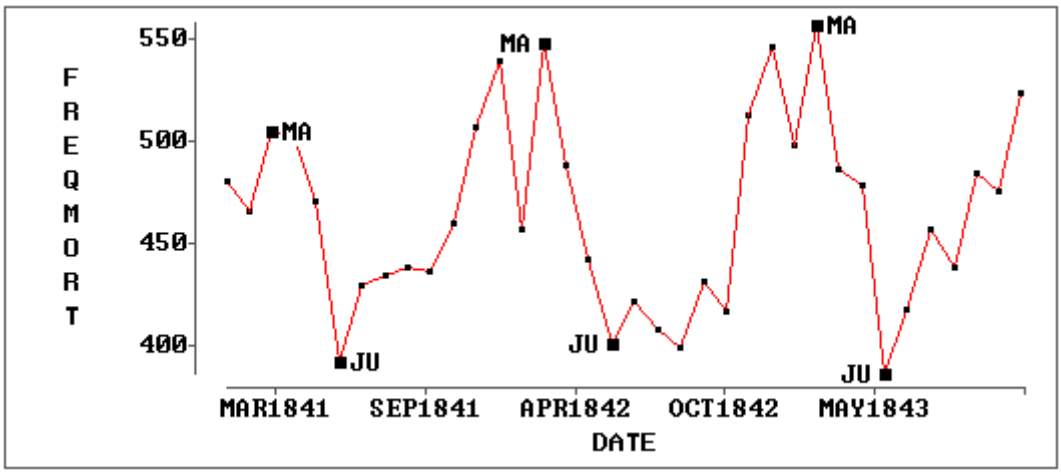
Guide de lecture de gauche à droite et de haut en bas

Cadre 1 : Les valeurs extrêmes du taux de natalité sont 10 et 53.
Cadre 2 : Répartition des pays suivant le taux de natalité (en ordonnées) et le taux de mortalité (en abscisses).
Cadre 3 : Répartiton des pays suivant le taux de natalité (en ordonnées) et le taux d'accroissement (en abscisses).
Cadre 4 (symétrique du cadre 2) : Répartition des pays suivant le taux de mortalité (en ordonnées) et le taux de natalité (en abscisses).
Cadre 5 : Les valeurs extrêmes du taux de mortalité sont 3 et 29.
Cadre 6 : Répartition des pays suivant le taux de mortalité (en ordonnées) et le taux d'accroissement (en abscisses).
Cadre 7 (symétrique du cadre 3) : Répartition des pays suivant le taux d'accroissement (en ordonnées) et le taux de natalité (en abscisses).
Cadre 8 (symétrique du cadre 6) : Répartition des pays suivant le taux d'accroissement (en ordonnées) et le taux de mortalité (en abscisses).
Cadre 9 : Les valeurs extrêmes du taux d'accroissement sont -0,2 et 4,0.

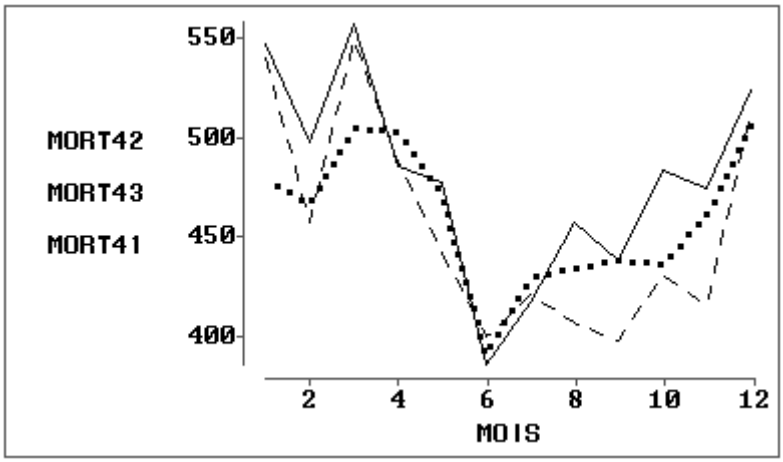
Le Line Plot

Le Line Plot, ou tracé de données reliées par des lignes, est adapté à la représentation des séries temporelles. L'utilisation de la couleur ou de tracés différents facilitera la visualisation en cas de prise en compte d'une troisième dimension déterminée par une variable nominale (cas du deuxième exemple ci-dessous).

Nombre d'enfants mort-nés en Belgique de janvier 1841 à décembre 1843



Nombre mensuel d'enfants mort-nés en Belgique, années 1841 à 1843

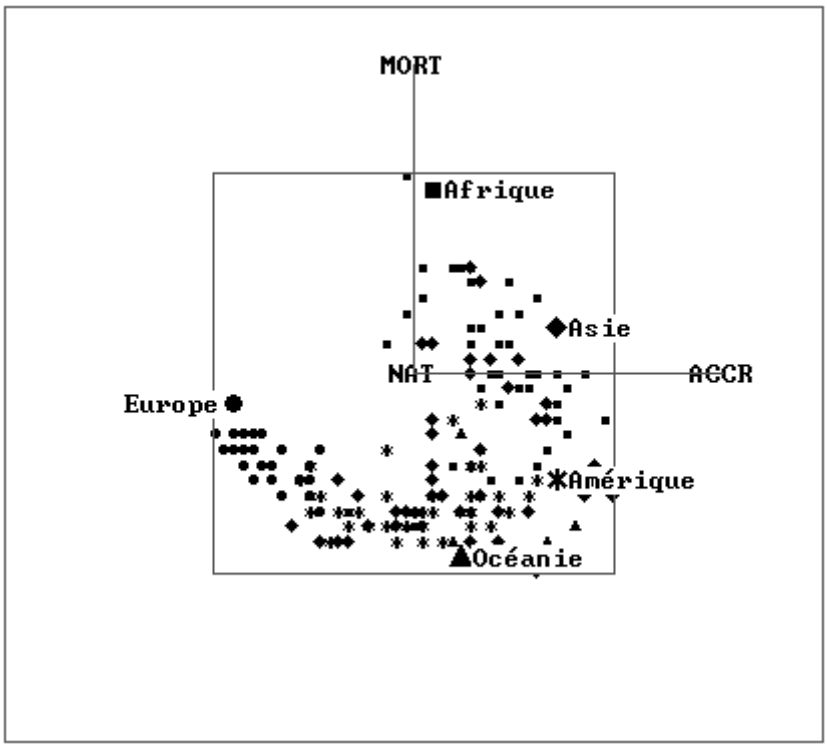


Le Rotating Plot

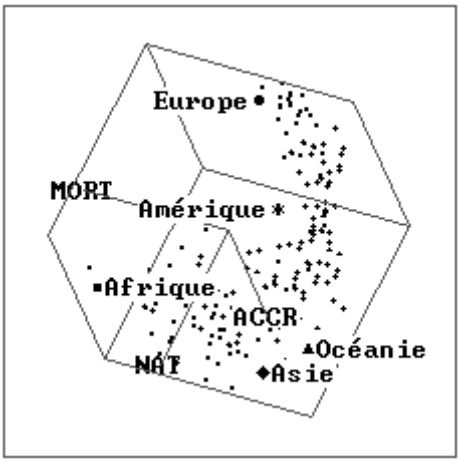
Le Rotating Plot, ou graphique dynamique de rotation en 3 dimensions, permet de repérer d'éventuelles structures qui ne sont ni visibles sur des graphiques statiques, ni détectables par des méthodes analytiques.

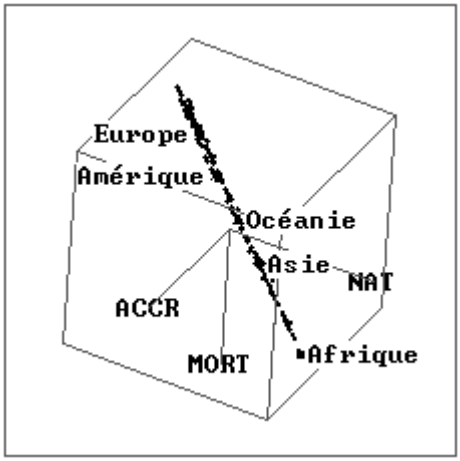
De ce point de vue, l'exemple ci-dessous pourrait paraître assez mal choisi, puisque l'on s'y intéresse à la répartition de nos 173 pays suivant le taux de natalité, le taux de mortalité et le taux d'accroissement naturel de la population.

Nota : L'axe NAT est ici perpendiculaire au plan de la page.



En faisant tourner le nuage de points, on met en évidence la relation liant les trois variables en question, ici ACCR = (NAT - MORT) /10 puisque les taux de natalité et de mortalité sont exprimés en pour mille et le taux d'accroissement en pour cent.

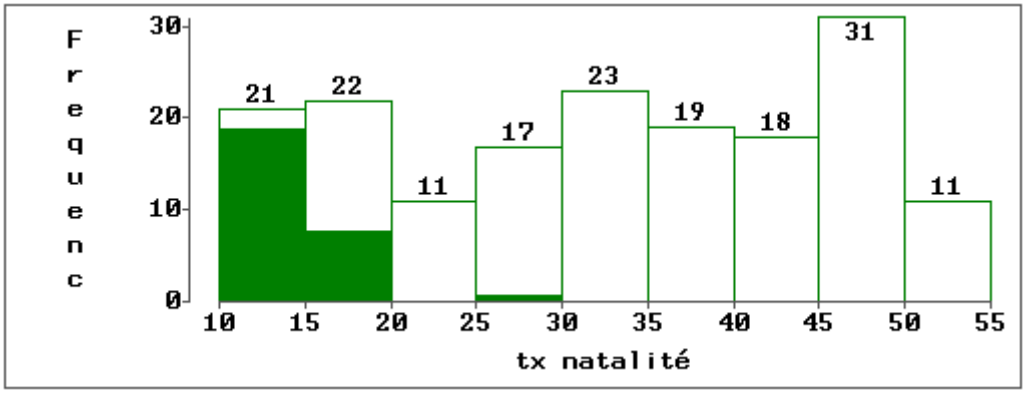


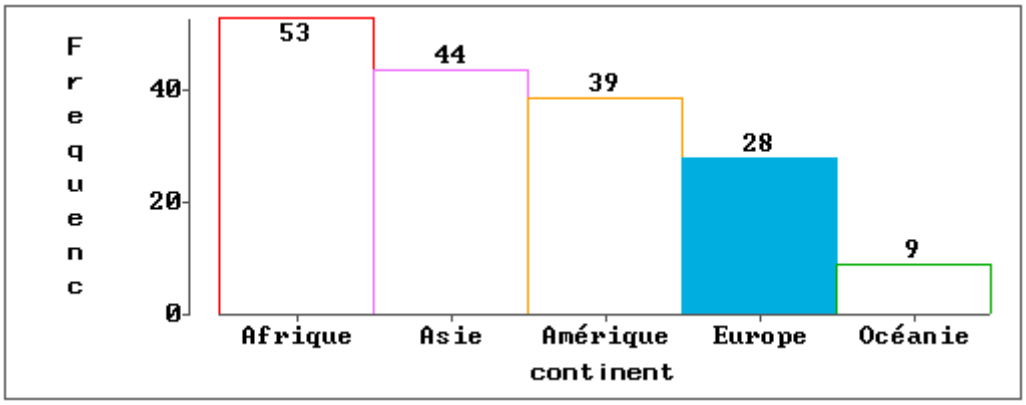


Interactivité

Il est tout à fait possible de regrouper plusieurs représentations graphiques dans une même fenêtre et d'animer l'ensemble grâce à l'interactivité.

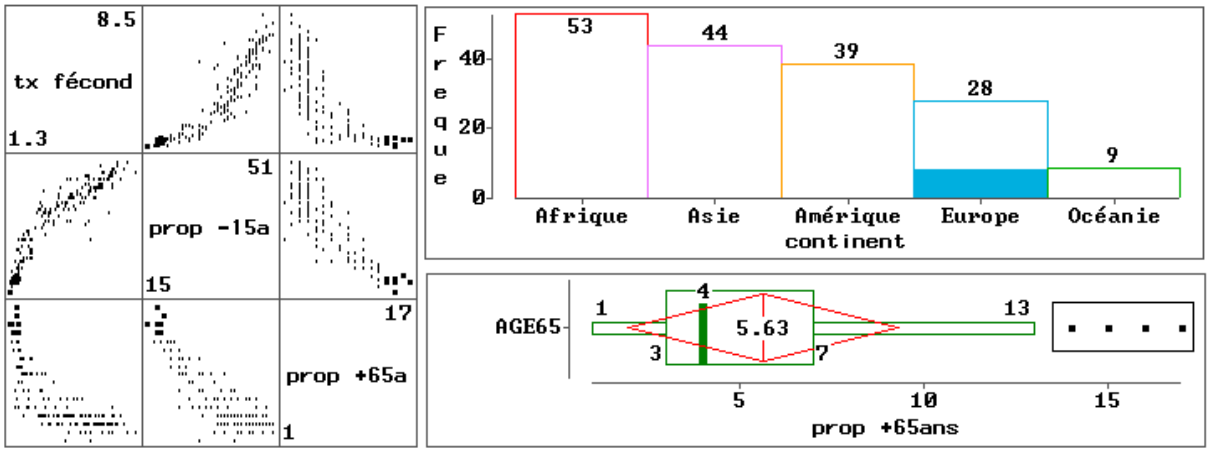
Regroupons par exemple dans une même fenêtre un histogramme relatif à la distribution du taux de natalité, et un Bar Chart donnant la répartition par continent des 173 pays observés. Si l'on sélectionne la barre Europe du Bar Chart, on verra alors instantanément se superposer à l'histogramme global un deuxième histogramme limité au cas de ce continent.





Plus spectaculaire encore, regroupons dans une même fenêtre ce même Bar Chart, le Box Plot relatif à la distribution de la proportion des plus de 65 ans ainsi qu'une matrice de diagrammes de dispersion mettant en oeuvre la proportion des moins de 15 ans, celle des plus de 65 ans et le taux de fécondité (nombre d'enfants par femme).

Si l'on sélectionne (en les encadrant) les quatre observations atypiques du Box Plot, on voit immédiatement se surimprimer la position des pays en question sur le Bar Chart et les diagrammes de dispersion (effet loupe) : ils sont situés en Europe, le taux de fécondité et la proportion des moins de 15 ans y sont particulièrement faibles, la proportion des plus de 65 ans y est particulièrement élevée.

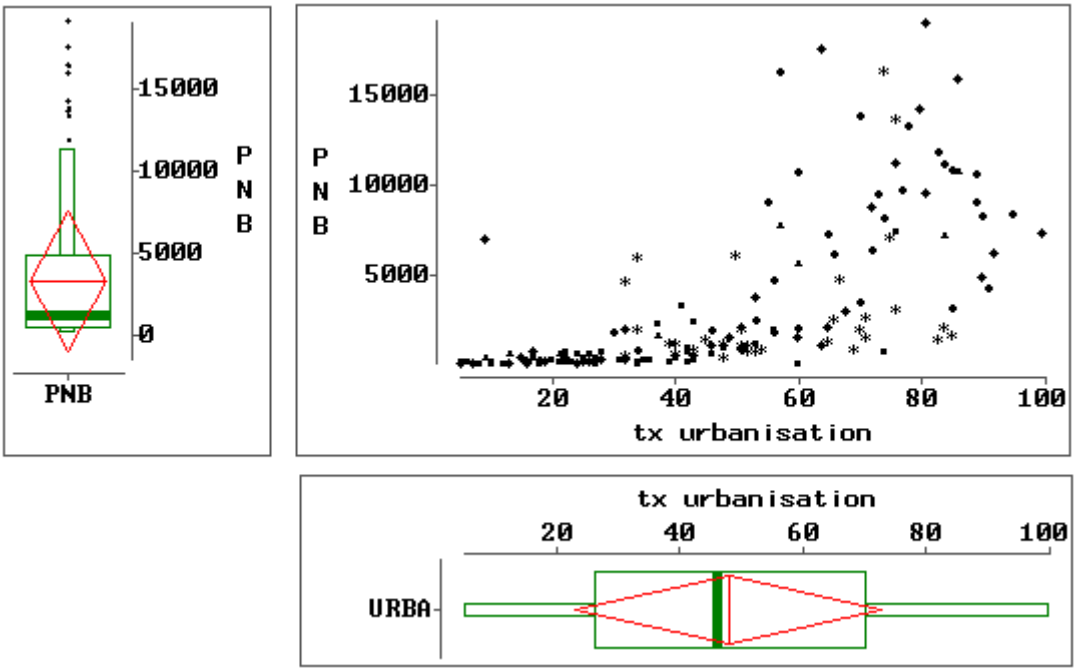


Réexpression des données

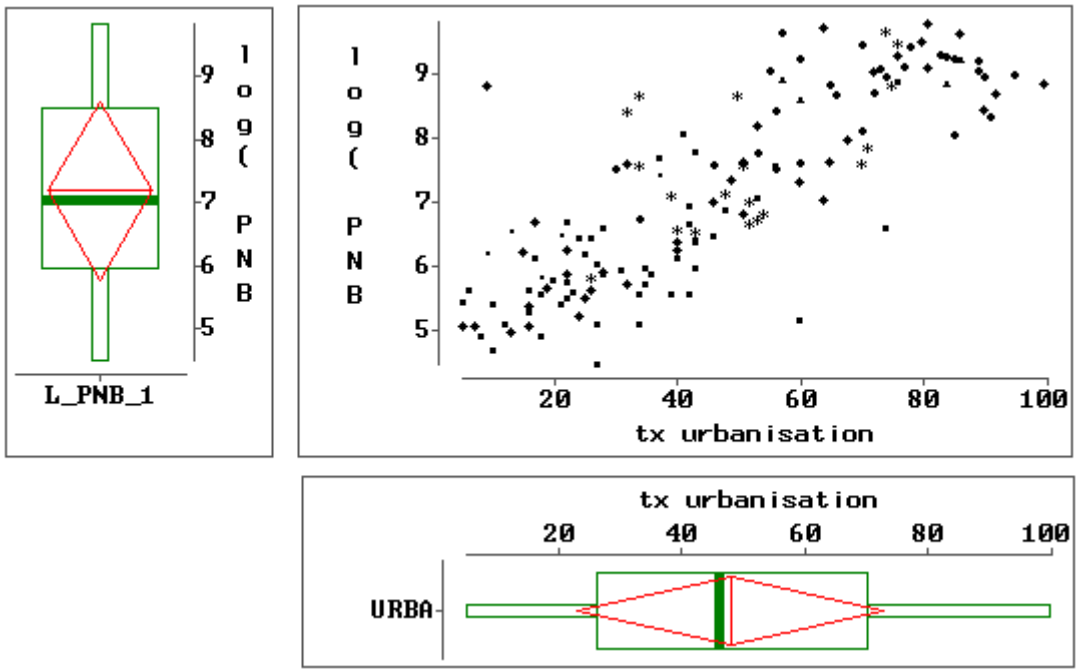
On l'a vu dans l'article « AED mode d'emploi », l'un des fondamentaux de l'analyse exploratoire des données est la transformation ou réexpression des variables, l'opération pouvant en particulier permettre de symétriser une distribution ou encore de linéariser une liaison.

Avec SAS/Insight, il est possible de directement transformer une variable à partir d'une représentation graphique, sans qu'il soit nécessaire de revenir au tableur.

Soit par exemple la fenêtre ci-dessous, dans laquelle nous avons regroupé deux Box Plots, le premier relatif à la distribution du PNB, le second à la distribution du taux d'urbanisation, et le Scatter Plot présentant la façon dont se répartissent les 173 pays observés suivant ces deux mêmes variables.



Sélectionnons le libellé de variable « PNB » sur le premier Box Plot ou le Scatter Plot, et transformons cette variable en son logarithme (logPNB) par l'intermédiaire du menu *Edit # Variables*. On obtient immédiatement la représentation transformée résultante, sur laquelle on va pouvoir constater, d'une part que la distribution de logPNB est quasi symétrique, d'autre part qu'il semble y avoir une liaison linéaire entre cette variable et le taux d'urbanisation.



Sophie DESTANDAU et Monique LE GUEN

Cet article a été publié dans :
DESTANDAU S., LADIRAY D., LE GUEN M., (1999), " *l'Analyse Exploratoire des données et SAS/INSIGHT*", Courrier des Statistiques, n°90, juin 1999, INSEE, pp3-44.